Neural Architecture Search of Echocardiography View Classifiers

Neda Azarmehr[a,*], Xujiong Ye[a], James P Howard[b], Elisabeth Sarah Lane[c], Robert Labs[c], Matthew J Shun-shin[b], Graham D Cole[b], Luc Bidaut[a], Darrel P Francis[b], Massoud Zolgharni[b,c]

[a]University of Lincoln, School of Computer Science, UK
[b]Imperial College London, National Heart and Lung Institute, UK
[c]University of West London, School of Computing and Engineering, UK

Abstract.

Purpose: Echocardiography is the most commonly used modality for assessing the heart in clinical practice. In an echocardiographic exam, an ultrasound probe samples the heart from different orientations and positions, thereby creating different viewpoints for assessing the cardiac function. The determination of the probe viewpoint forms an essential step in automatic echocardiographic image analysis.

Approach: In this study, convolutional neural networks are used for the automated identification of 14 different anatomical echocardiographic views (larger than any previous study) in a dataset of 8,732 videos acquired from 374 patients. Differentiable architecture search approach was utilised to design small neural network architectures for rapid inference while maintaining high accuracy. The impact of the image quality and resolution, size of the training dataset, and number of echocardiographic view classes on the efficacy of the models were also investigated.

Results: In contrast to the deeper classification architectures, the proposed models had significantly lower number of trainable parameters (up to 99.9% reduction), achieved comparable classification performance (accuracy 88.4-96.0%, precision 87.8-95.2%, recall 87.1-95.1%) and real-time performance with inference time per image of 3.6-12.6ms.

Conclusion: Compared with the standard classification neural network architectures, the proposed models are faster and achieve comparable classification performance. They also require less training data. Such models can be used for real-time detection of the standard views.

Keywords: Deep Learning, Echocardiography, Neural Architecture Search, View Classification, AutoML.

***Neda Azarmehr, n.azarmehr@gmail.com**

1 Introduction

Echocardiography or cardiac ultrasound imaging is the modality of choice for the diagnosis of cardiac pathology. Echocardiographic (echo) measurements provide quantitative diagnostic markers of cardiac function. Portability, speed, and affordability are the advantages of echo.

Echo examinations are typically focused upon protocols containing diverse probe positions and orientations providing several views of the heart anatomy. Standard echo views require imaging the heart from multiple windows. Each window is specified by the transducer position and includes

parasternal, apical, subcostal and suprasternal. The orientation of the echo imaging plane produces views such as long axis, short axis, four-chamber, and five-chamber.[1]

Interpretation of echo images begins with view detection. This is a time-consuming and manual process that requires specialised training and is prone to inter- and intra-observer variability. Echo images are very similar and can be particularly challenging for an operator to successfully categorise.

Therefore, accurate automatic classification of heart views has several potential clinical applications such as improving workflow, guiding inexperienced users, reducing inter-user discrepancy, and improving accuracy for high throughput of echo data and subsequent diagnosis.

In most current clinical practice, images from different modalities are managed and stored in Picture Archiving and Communication Systems (PACS). Recently, add-on echo software packages, such as EchoPAC (GE Healthcare) and QLAB (Philips), attempt to automate the analysis and diagnosis process. However, they still necessitate human involvement in detecting relevant views. As previously stated, echocardiography image frames are not easily discernible by the operator, plus there is often background noise. Therefore, automatic view classification could be widely beneficial for pre-labelling large datasets of unclassified images.[2,3]

Application of machine learning algorithms in computer vision has improved the accuracy and time-efficiency of automated image analysis, particularly automated interpretation of medical images.[4–7] However, traditional machine learning methods are constructed using complex processes and tend to have a restricted scope and effectiveness.[8,9] Recent advances in the design and application of deep neural networks have resulted in increased possibilities when automating medical image-based diagnosis.[10,11]

1.1 Approaches to neural network design

Convolutional neural networks (CNNs) are extremely effective at learning patterns and features from digital images and have demonstrated success in many image classification tasks.[12,13] However, this success has been accompanied by a growing demand for architecture engineering of increasingly more complex deep neural networks through a time-consuming and arduous manual process. Moreover, the developed architectures are usually dependent on the particular image dataset used in the design process, and adapting the architectures to new datasets remains a very difficult task that relies on extensive trial and error process and expert knowledge.

Recently, increased attention has been paid to emerging algorithmic solutions, such as Neural Architecture Search (NAS), to automate the manual process of architecture design, and these have accomplished highly competitive performance in image classification tasks.[14–17] NAS can actually be considered as a subfield of automated machine learning (AutoML).[18]

Pivotal to the NAS architecture is the creation of a large collection of potential network architectures. These options are subsequently explored to determine an ideal output with a specific combination of training data and constraints, such as network size. Initial NAS approaches, such as reinforcement learning[19,20] and evolution,[21] search for complete network topology, thus involving extremely large search spaces comprised of arbitrary connections and operations between neural network nodes. Such complexity results in using massive amounts of energy and requiring thousands of GPU hours or million-dollar cloud compute bills[22] to design neural network architectures.

Successful NAS approaches, such as Efficient Neural Architecture Search (ENAS) from Google Brain[15] and more recently Differentiable Architecture Search (DARTS),[16] have been shown to reduce the search costs by orders of magnitude, requiring $\sim$100x fewer GPU hours. These methods

leverage an important observation that popular CNN architectures often contain repeating blocks or are stacked sequentially. Their effectiveness is thus owing to the key idea of focusing on finding a small optimal computational cell (as the building block of the final architecture), rather than searching for a complete network. The size of the search space is therefore significantly reduced since the computational cells contain considerably fewer layers than the whole network architecture, which would make such approaches potentially viable for solving real-world challenges.

The DARTS method has been shown to outperform ENAS in terms of the GPU hours required for the search process.[16] While most NAS studies report experimental results using standard image datasets such as CIFAR and ImageNet, the effectiveness of DARTS on scientific datasets, including medical images, has also been demonstrated. In this study, the DARTS method for designing customised architectures has been adopted.

1.2 Related work on echocardiography view classification

Most previous studies on automatic classification of echocardiographic views have used handcrafted features and traditional machine learning techniques, achieving varying degrees of success in classifying a limited number of common echocardiographic views.[22–30] Following the recent success of deep convolutional neural networks in computer vision, and particularly for image classification tasks, there has been a handful of reports on the application of deep learning for cardiac ultrasound view detection. Herein, we have focused on such studies.

Gao et al.[30] proposed a fused CNN architecture by integrating a deep learning network along the spatial direction, and a hand-engineered feature network along the temporal dimension. The final classification result for the two-strand-network was obtained through a linear combination of the classification scores obtained from each network. They used a dataset of 432 image sequences

acquired from 93 patients. For each strand of CNN network implemented using Matlab, it took 2 days to process all images. Their model achieved an average accuracy rate of 92.1% when classifying 8 different echocardiographic views.

In another study,[31] view identification formed part of an automated pipeline designed for the interpretation of echocardiograms. The standard VGG architecture was employed as the CNN model, and 6 different echocardiographic views were included in the study. The class label for each video was assigned by taking the majority decision of predicted view labels on the 10 frames extracted from the video. The overall classification accuracy, calculated from the reported confusion matrix, was 97.7%, and no results for single image classification was reported. In a follow-up study,[3] they included 23 views (9 of which were 3 apical planes, each one divided into 'no occlusions', 'occluded LA', and 'occluded LV' categories) from 277 echocardiograms. The reported overall accuracy of the VGG model dropped to 84% at an individual image level, with the greatest challenge being distinctions among the various apical views. By averaging across multiple images from each video, higher accuracies could be achieved.

Madani et al.[32] proposed a CNN model to classify 12 standard B-mode echocardiographic views (15 views, including Doppler modalities) using a dataset of 267 transthoracic studies (90% used for training-validation, and 10% for testing). An inference latency of 21ms per image was achieved for images with a size of 60×80 pixels. They also reported an average overall accuracy of 91.7% for classifying single frames, compared to an average of 79.4% for expert echocardiographers classifying a subset of the same test images. However, this may not be a fair comparison as the expert humans were given the same downsampled images that were fed into the CNN model, but the human experts are not trained and have no experience of working with such low-resolution images. Later on, they reported an improved classification accuracy of 93.64% by first applying

a segmentation stage, where the field of view was extracted from the images using U-net model[33] and the isolated image segment was then fed into the classifier.[34]

In a more recent study,[6] a CNN model was proposed with the aim to balance accuracy and effectiveness. The design was inspired by the Inception[35] and DenseNet[36] architectures. The performance of the model was examined using a dataset of 2559 image sequences from 265 patients, and an overall accuracy of 98.3% was observed for classifying 7 echocardiographic views. The reported inference time was 4.4 ms and 15.9 ms when running the model on the GPU and CPU, respectively, for images with a size of 128×128 pixels.

Vaseli et al.[37] reported on designing a lightweight model with the knowledge of three state-of-the-art networks (VGG16, DenseNet, and ResNet) for classifying 12 echocardiographic views. A maximum accuracy of 88.1% was observed using their lightweight models, with a minimum inference time of 52μs for images with a size of 80×80 pixels. However, the reported accuracies are provided for classifying cine loops, and are computed as the average of the predictions for all constituent frames in each cine loop. It is unclear how many frames constituted a cine loop. For a cine loop containing 120 frames (time-window of 2s acquired at 60 frames/s), therefore, an inference time of $\geq$6.2ms would be required to achieve the reported accuracy. A more rigorous examination of their models also seems necessary and, as apparent from the provided confusion matrices, a great majority of the reported misclassifications, seen as a failure of the models, occurred for parasternal short-axis views.

1.3 Main contributions

Given our two competing objectives of minimising the neural network size and maximising its prediction accuracy, this study aims to adopt the recent NAS solution of DARTS for designing

efficient neural networks. To the best of our knowledge, no other study has applied DARTS to the complex problem of echocardiographic views classification.

In our study, we also aimed at including subclasses of a given echocardiographic view. In general, the more numerous the view classes, the more difficult the task of distinguishing the views for the CNN model. This is because if a group of images is considered as a single view in one study and as multiple views in another, those multiple views are likely to be relatively similar in appearance. Perhaps this is one of the primary reasons for the wide range of accuracies (84-97%) reported in the literature.

We have previously reported on preparation and annotation of a large patient dataset, covering a range of pathologies and including 14 different echocardiographic views, which we used for evaluating the performance of existing standard CNN architectures.[38] In this study, we will use this dataset to design customised network architectures for the task of echo view classification.

The input image resolution could potentially impact the classification performance. In case of aggressively downsampled images, the relevant features may in fact be lost, thus lowering the classification accuracy. On the other hand, unnecessarily large images would result in more computations. Nevertheless, all previous reports considered one particular (but dissimilar in different studies) image resolution, the selection of which was always unexplained. Herein, we have thus looked at the impact of different input image resolutions.

The accuracy of deep learning classifiers is largely dependent on the size of high-quality initial training datasets. Collecting an adequate training dataset is often the primary obstacle of many computer vision classification tasks. This could be particularly challenging in medical imaging where the size of training datasets are scarce, e.g. because the images can only be annotated by skilled experts. Hence, it would be advantageous to require less training data. Therefore, we

examined the influence of the size of training data on the model's performance for each of the investigated networks in this study.

No matter how ingenious the deep learning model, image quality places a ceiling on the reliability of any automated image analysis. Echocardiograms inherently suffer from relatively poor image quality. Therefore, we also looked at the impact of image quality on the classification performance.

In light of the above, the main contributions of this study can be summarised as follows:

- Inclusion of 14 different anatomical echocardiographic views (outlined in Figure 1); larger than any previous study. We also examined the cases when only 7 or 5 different views were included to investigate the impact of the number of views on the detection accuracy.

- Analysis of three well-known network topologies and of a proposed neural network, obtained from applying NAS techniques to design network topologies with far fewer trainable parameters and comparable/better accuracy for echo view classification.

- Analysis of computational and accuracy performance of the developed models using our large-scale test dataset.

- Analysis of the impact of the input image resolution; 4 different image sizes were investigated.

- Analysis of the influence of the size of training data on the model's performance for all investigated networks.

- Analysis of the correlation between the image quality and accuracy of the model for view detection.

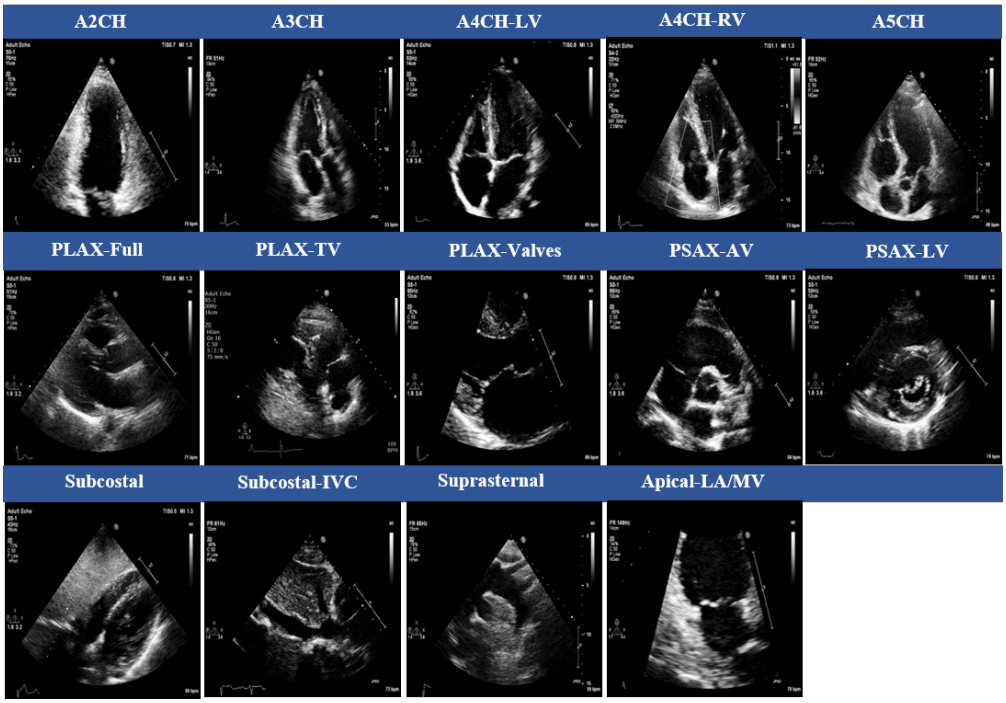

Fig 1 The 14 cardiac views in transthoracic echocardiography: apical two-chamber (A2CH), apical three-chamber (A3CH), apical four-chamber left ventricle focused (A4CH-LV), apical four-chamber right ventricle focused (A4CH-RV), apical five-chamber (A5CH), parasternal long-axis (PLAX-Full), parasternal long-axis tricuspid valve focused (PLAX-TV), parasternal long-axis valves focused (PLAX-Valves), parasternal short-axis aortic valve focused (PSAX-AV), parasternal short-axis left ventricle focused (PSAX-LV), subcostal (Subcostal), subcostal view of the inferior vena cava (Subcostal-IVC), suprasternal (Suprasternal), and apical left atrium mitral valve focused (LA/MV).

2 Dataset

In this section, a brief account of the patient dataset used in this study is provided. A detailed description, including patient characteristics, can be found in Howard et al.[38]

A random sample of 374 echocardiographic examinations of different patients and performed between 2010 and 2020 was extracted from Imperial College Healthcare NHS Trust's echocardiogram database. The acquisition of the images was performed by experienced echocardiographers and according to standard protocols, using ultrasound equipment from GE and Philips manufacturers.

Ethical approval was obtained from the Health Regulatory Agency (Integrated Research Ap-

plication System identifier 243023). Only studies with full patient demographic data and without intravenous contrast administration were included. Automated anonymization was performed to remove all patient-identifiable information.

The videos were annotated manually by an expert cardiologist (JPH), categorising each video into one of 14 classes which are outlined in Figure 1. Videos thought to show no identifiable echocardiographic features, or which depicted more than one view, were excluded. Altogether, this resulted in 9,098 echocardiographic videos. Of these, 8,732 (96.0%) videos could be classified as one of the 14 views by the human expert. The remaining 366 videos were not classifiable as a single view, either because the view changed during the video loop, or because the images were completely unrecognisable. The cardiologist's annotations of the videos were used as the ground truth for all constituent frames of that video.

DICOM-formatted videos of varying image sizes (480×640, 600×800, and 768×1024 pixels) were then split into constituent frames, and three frames were randomly selected from each video

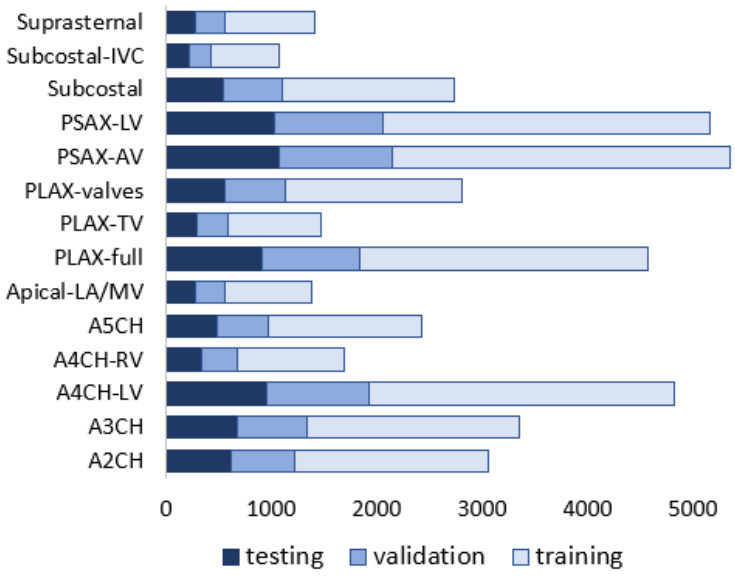

Fig 2 Distribution of data in the training, validation and test dataset; values show the number of frames in a given class.

to represent arbitrary stages of the heart cycle, resulting in 41,321 images. The dataset was then randomly split into training (24791 images), validation (8265 images), and testing (8265 images) sub-datasets in a 60:20:20 ratio. Each sub-datasets contained frames from separate echo studies to maintain sample independence.

The relative distribution of echo view classes labelled by the expert cardiologist is displayed in Figure 2 and indicates an imbalanced dataset, with a ratio of 3% (Subcostal-IVC view as the least represented class) to 13% (PSAX-AV view as the dominant view).

3 Method

Details of the well-known classification network architectures investigated in this study (i.e., VGG16, ResNet18, and DenseNet201) can be found in relevant resources.[36,39,40] Here, a detailed description of the designed CNN models will be provided.

3.1 DARTS method

Proposed by Liu et al. in 2019,[16] DARTS formulates the architecture search task in a differentiable manner. Unlike conventional approaches of applying evolution[21,41] or reinforcement learning[14,42] over a discrete and non-differentiable search space, DARTS is based on the continuous relaxation of the architecture representation, allowing an efficient search of the architecture using gradient descent.

DARTS method consists of two stages: architecture search and architecture evaluation. Given the input images, it first embarks on an architecture search to explore for a computation cell (a small unit of convolutional layers) as the building block of the neural network architecture. After the architecture search phase is complete and the optimal cell is obtained based on its validation

performance, the final architecture could be formed from one cell or a sequential stack of cells. The weights of the optimal cell learnt during the search stage are then discarded, and are initialised randomly for retraining the generated neural network model from scratch.

A cell, depicted in Figure 3, is an ordered sequence of several nodes in which one or multiple edges meet. Each node $C^{(i)}$ represents a feature map in convolutional networks. Each edge (i,j) is associated with some operation $O^{(i,j)}$, transforming the node $C^{(i)}$ to $C^{(j)}$. This could be a combination of several operations, such as convolution, max-pooling, and ReLU.

Each intermediate node $C^{(j)}$ is computed based on all of its predecessors as:

$$C^{(j)} = \sum_{i<j} O^{(i,j)}\left(C^{(i)}\right) \tag{1}$$

Instead of applying a single operation (e.g., 5×5 convolution), and evaluating all possible operations independently (each trained from scratch), DARTS places all candidate operations on each edge (e.g., 5×5 convolution, 3×3 convolution, and max-pooling represented in Figure 3 by red, blue, and green lines, respectively). This allows sharing and training their weights in a single process. The task of learning the optimal cell is effectively finding the optimal placement of operations at the edges.

The actual operation at each edge is then a linear combination of all candidate operations O(i,j), weighted by the softmax output of the architecture parameters $\alpha^{(i,j)}$:

$$\bar{O}^{(i,j)}(C) = \sum_{o\in\partial} \frac{exp({\alpha_o}^{(i,j)})}{\sum_{o'\in\partial} exp(\alpha_{o'}^{(i,j)})} O(C) \tag{2}$$

Optimization of the continuous architecture parameters α is carried out using gradient descent

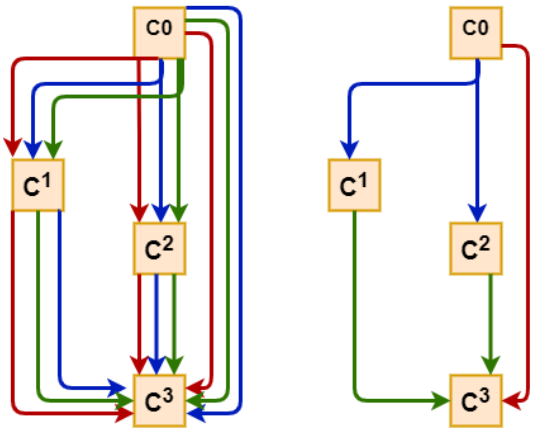

Fig 3 Schematic of a DARTS cell. Left: a computational cell with four nodes C^0-C^3. Edges connecting the nodes represent some candidate operations (e.g., 5×5 convolution, 3×3 convolution, and max-pooling represented in Figure 3 by red, blue, and green lines, respectively). Right: the best-performing cell learnt from retaining the optimal operations. Figure inspired by Elsken et al.[43]

on the validation loss. The mixed operation $\bar{O}^{(i,j)}$ is then replaced by the operation $O^{(i,j)}$ corresponding to the highest weight:

$$O^{(i,j)} = argmax_{o \in \partial} \quad \alpha_0^{(i,j)} \tag{3}$$

An example final cell architrave is displayed in the right panel, in Figure 3. The task of architecture search is learning a set of continuous variables in vector $\alpha^{(i,j)}$.

The training loss $\mathcal{L}_{train}$ and validation loss $\mathcal{L}_{val}$ are determined by the architecture parameters α and the weights ω in the network. The learning of α is performed in conjunction with learning of ω within all the candidate operations (e.g., weights of the convolution filters).

DARTS seeks to find the architecture α^* that minimises $\mathcal{L}_{val}(\omega^*, \alpha^*)$, where the weights ω^* associated with the architecture minimise the training loss $\omega^* = \text{argmin}_\omega \mathcal{L}_{train}(\omega, \alpha^*)$, This indi-

cates a bi-level optimization problem as:

$$\min_{\alpha} \quad \mathcal{L}_{val}(\omega^*(\alpha), \alpha) \tag{4}$$

$$such.that \quad \omega^*(\alpha) = argmin_{\omega} \quad \mathcal{L}_{train}(\omega, \alpha) \tag{5}$$

It is computationally expensive to solve the optimization problem precisely; i.e., computing the true loss by training ω for each architecture. Utilising a one-step approximation, the training of α and ω is performed by alternating the gradient steps in the weights and the architecture parameters.

The weights are optimized by descending in the direction $\nabla_{\omega}\mathcal{L}_{train}(\omega, \alpha)$, while α is optimized by descending in the direction $\nabla_{\alpha}\mathcal{L}_{val}(\omega$ - $\xi\nabla_{\omega}\mathcal{L}_{train}(\omega, \alpha),\alpha)$, where ξ is equal to the learning rate for the weights optimiser.

Two types of cells are defined and optimized in DARTS:

- Normal Cell which maintains the output spatial dimension the same as input
- Reduction Cell which reduces the output spatial dimension while increasing the number of filters/channels

The final architecture is then formed by stacking these cells.

3.2 DARTS parameters for architecture search

For the stage of architecture search, 80% of the dataset was held out for equally-sized training and validation subsets, and 20% for testing. Images were normalised and downsampled to 4 different

sizes of 32×32, 64×64, 96×96, and 128×128 pixels, with corresponding batch sizes of 64, 14, 8, and 4, respectively.

The following candidate operations were included in the architecture search stage: 3×3 and 5×5 separable convolutions, 3×3 and 5×5 dilated separable convolutions, 3×3 max-pooling, 3×3 average-pooling, skip-connection, and zero. For the convolutional operations, a ReLU-Conv-BN order was used. If applicable, the operations were of stride one. The convolved feature maps were padded to preserve their spatial size.

A network of 8 cells was then used to conduct the search for a maximum of 30 epochs. The initial number of channels was 16 to make sure the network could fit into a single GPU. Stochastic Gradient Decent (SGD) with a momentum of 0.9, initial learning rate of 0.1, and weight decay of 3×10^{-4} was used to optimise the weights. To obtain enough learning signal, DARTS utilises zero initialization for architecture variables indicating the same amount of attention over all possible operations as it is taking the softmax after each operation.

Adam optimiser[44] with an initial learning rate of 0.1, momentum of (0.5, 0.999), and weight decay of 10^{-3} were used as the optimiser for α.

3.3 Models training parameters

Training occurred subsequently, using annotations provided by the expert cardiologist. It was carried out independently for each of the 4 different image sizes of 32×32, 64×64, 96×96, and 128×128 pixels. Identical training, validation, and testing datasets were used in all network models. The validation dataset was used for early stopping to avoid redundant training and overfitting. Each model was trained until the validation loss plateaued. The test dataset was used for the performance assessment of the final trained models. The DARTS models were kept blind to the test

dataset during the stage of architecture search.

Adam optimiser with a learning rate of 10^{-4} and a maximum number of 800 epochs was used for training the models. The cross-entropy loss was used as the networks objective function. For training the DARTS model, a learning rate of 0.1 deemed to be a better compromise between speed of learning and precision of result and was therefore used. A batch size of 64 or the maximum which could be fitted on the GPU (if $<$64) was employed.

It is evident from Figure 2 that the dataset is fairly imbalanced with unequal distribution of different echo views. To prevent potential biases towards more dominant classes, we used online batch selection where the equal number of samples from each view were randomly drawn (by over-sampling of underrepresented classes). This led to training on a balanced dataset representing all classes in every epoch. An epoch was still defined as the number of iterations required for the network to meet all images in the training dataset.

3.4 Evaluation metrics

Several metrics were employed to evaluate the performance of the investigated models in this study. Overall accuracy was calculated as the number of correctly classified images as a fraction of the total number of images. Macro average precision and recall (average overall views of per-view measures) were also computed. F1 score was calculated as the harmonic mean of the precision and recall. Since this study is a multi-class problem, F1 score was the weighted average, where the weight of each class was the number of samples from that class.

PyTorch[45] was used to implement the models. For the computationally intensive stage of architecture search, a GPU server equipped with 4 NVIDIA TITAN RTX GPUs with 64 GB of memory was rented. For the subsequent training of the searched networks and also the standard models, the

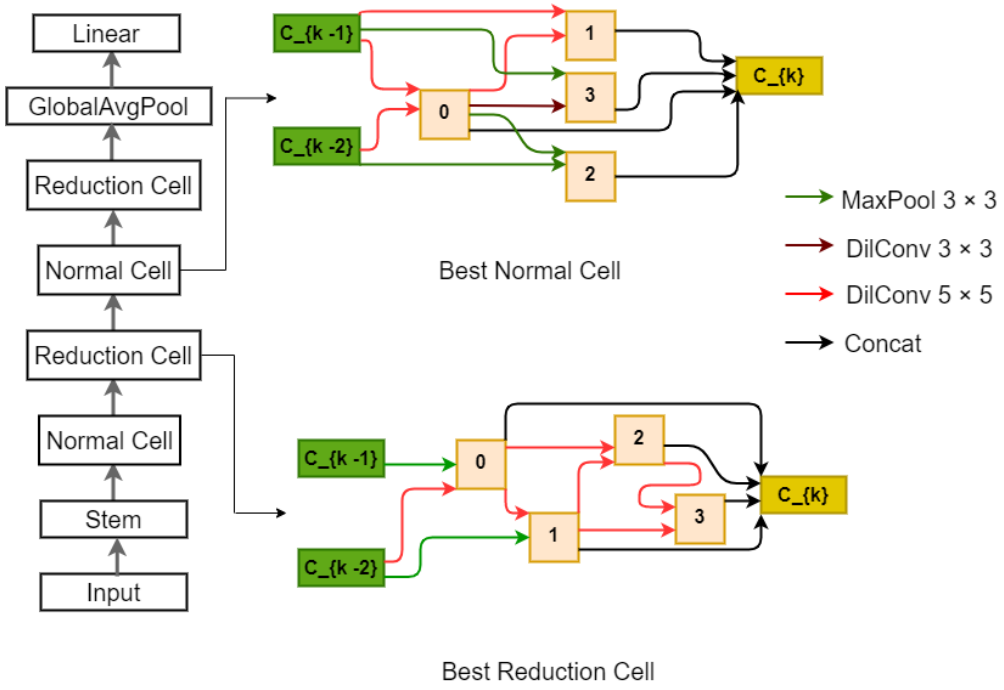

Fig 4 Optimal normal and reduction cells for the input image size of 128×128 pixels, as suggested by the DARTS method, where 3×3 and 5×5 dilated separable convolutions, 3×3 max-pooling, and skip-connection operations have been retained from the candidate operations initially included. Each cell has 2 inputs which are the cell outputs in the previous two layers. The output of the cell is defined as the depth-wise concatenation of all nodes in the cell. A schematic view of the "2-cell-DARTS", formed from a sequential stack of 2 cells, is also displayed on the left. Stem layer incorporates a convolution layer and a batch normalisation layer.

utilised GPU was an Nvidia QUADRO M5000 with 8 GB of memory, representing a more widely accessible hardware for real-time applications. Inference time (latency time for classifying each image) was also estimated with the trained models running on the GPU. To this end, a total of 100 images were processed in a loop, and the average time was recorded. All training/prediction computations were carried using identical hardware and software resources, allowing for a fair comparison of computational time-efficiency between all network models investigated in this study.

The number of trainable parameters in the model, as well as the training time per epoch was also recorded for all CNN networks.

Table 1 Experimental results on the test dataset for input sizes of (32×32), (64×64), (96×96) and (128×128) and different network topologies. Accuracy is ratio of correctly classified images to the total number of images; precision and recall are the macro average measures (average overall views of per-view measures); F1 score is the harmonic mean of precision and recall. The values in bold indicate the best performance for each measure.* For these experiments, a maximum batch size of <64 could be fitted on the GPU.

Network	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Parameters (thousands)	Inference Time (ms)	Time/epoch (s)
				(32×32)			
1-cell-DARTS	88.4	87.8	87.1	87.4	**58**	**3.6**	**41**
2-cell-DARTS	**93.0**	**92.5**	**92.3**	**92.3**	411	7.0	46
ResNet18	90.6	89.9	89.7	89.8	11,177	11.8	184
Vgg16	90.7	89.9	89.5	89.6	134,316	8.3	210
DenseNet201	88.3	87.9	87.0	87.4	20,013	119	1303
				(64×64)			
1-cell-DARTS	90.0	89.4	88.7	89.0	**92**	**6.5**	**81**
2-cell-DARTS	**95.0**	**94.7**	**94.2**	**94.4**	567	12.6	121
ResNet18	92.1	91.5	91.7	91.5		12.0	185
Vgg16	92.4	91.5	92.2	91.8		8.5	240
DenseNet201	93.1	92.5	92.8	92.6		127.3	1322
				(96×96)			
1-cell-DARTS	93.2	92.8	92.3	92.5	**101**	**7.2**	**141**
2-cell-DARTS	**95.4**	**95.1**	**94.9**	**94.9**	669	14.2	264
ResNet18	93.1	92.4	92.2	92.3		12.1	186
Vgg16	93.6	92.9	93.0	92.9		8.6	276
DenseNet201	93.8	93.0	93.3	93.1		129.0	1336
				(128×128)			
1-cell-DARTS	92.5	92.3	91.4	91.8	**89**	**5.9**	**180**
2-cell-DARTS	**96.0**	**95.2**	**95.1**	**95.1**	545	11.8	380*
ResNet18	92.9	92.6	92.2	92.4		12.2	196
Vgg16	93.2	92.1	92.7	92.3		9.0	429*
DenseNet201	93.8	93.1	93.2	93.1		129.4	1605*

4 Results and Discussion

4.1 Architecture search

The search took ∼6, 23, 42, and 92 hours for image sizes of 32×32, 64×64, 96×96, and 128×128 pixels, respectively, on the computing infrastructure described earlier (section 3.4). Figure 4 displays the best convolutional normal and reduction cells obtained for the input image size of 128×128 pixels. The retained operations were 3×3 and 5×5 dilated convolutions, 3×3 max-pooling, and skip-connection. Each cell is assumed to have 2 inputs which are the outputs from the previous and penultimate cells. The output of the cell is defined as the depth-wise concatenation of all nodes in the cell.

Two network architectures were assembled from the optimal cell; "1-cell-DARTS" comprised of one cell only, and "2-cell-DARTS" formed from a sequential stack of 2 cells. Addition of more cells to the network architecture did not significantly improve the prediction accuracy, as reported in the next section, but increased the number of trainable parameters in the model and thus the inference time for view classification. Therefore, the models with more than 2 cells, i.e. architectures with redundancy, were judged as being comparatively inefficient and thus discarded. Figure 4 (left side) also displays the full architecture for the "2-cell-DARTS" model for the input image size of 128×128 pixels.

4.2 View classification

Results for 5 different network topologies and different image sizes are provided in Table 1. Despite having significantly fewer trainable parameters, the two DARTS models showed competitive results when compared with the standard classification architectures (i.e., VGG16, ResNet18, and DenseNet201). The 2-cell-DARTS model, with only ∼0.5m trainable parameters, achieves the

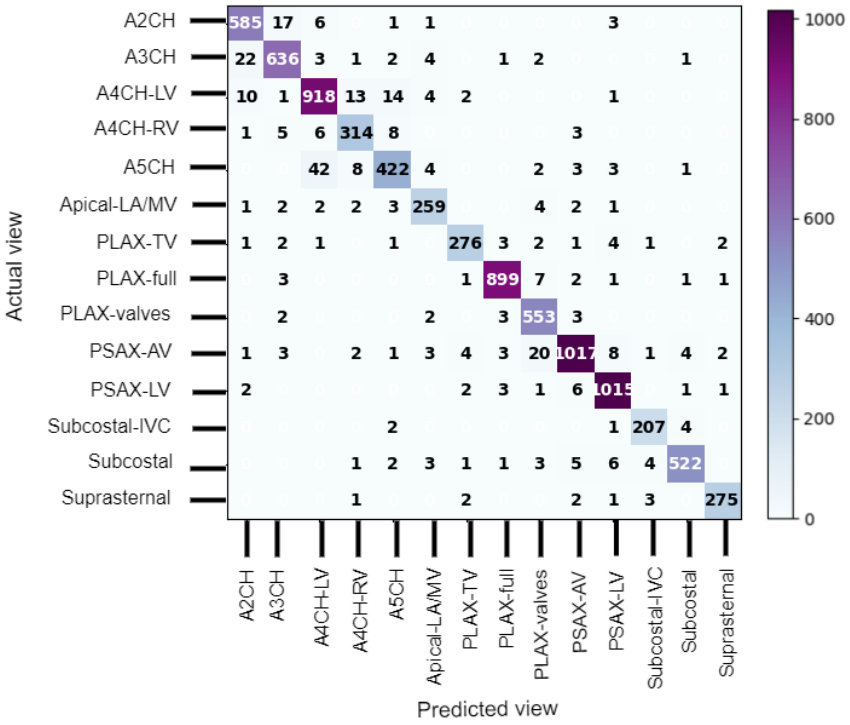

Fig 5 Confusion matrix for the 2-cell-DARTS model and input image resolution of 128×128 pixels.

best accuracy (93-96%), precision (92.5-95.2%), and recall (92.3-95.1%) among all networks and across all input image resolutions. Deeper standard neural networks, if employed for echo view detection, would therefore be significantly redundant, with up to 99% redundancy in trainable parameters.

On the other hand, while maintaining a comparable accuracy to standard network topologies, the 1-cell-DARTS model has $\leq$0.09m trainable parameters and the lowest inference time amongst all models and across different image resolutions (range 3.6-7.2ms). This would allow processing about 140-280 frames per second, thus making real-time echo view classification feasible.

Compared with manual decision making, this is a significant speedup. Although the identification of the echo view by human operators is almost instantaneous (at least for easy cases), the average time for the overall process of displaying/identifying/recording the echo view takes several seconds.

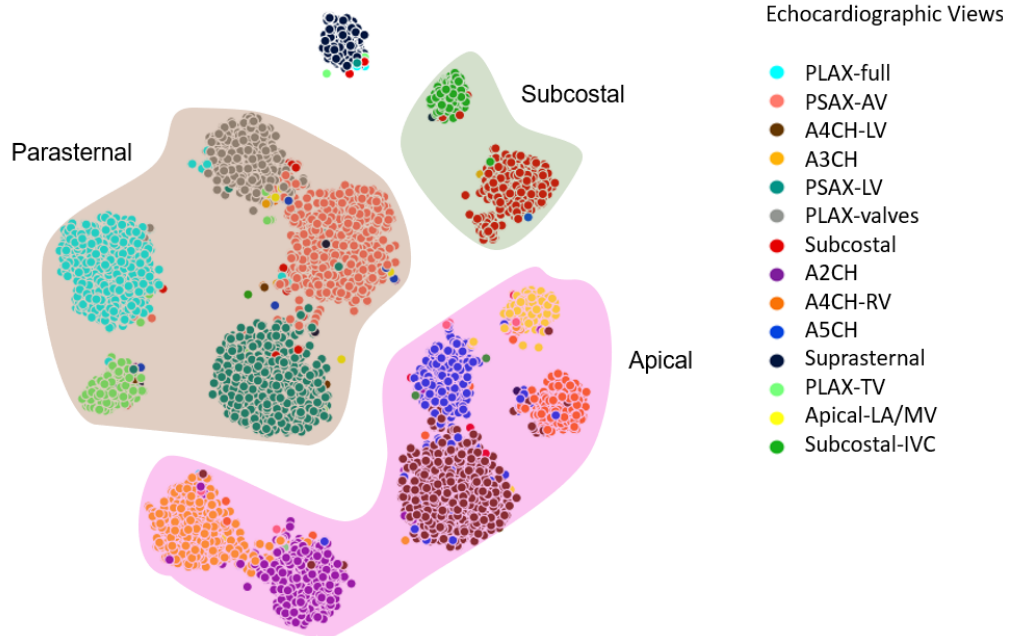

Fig 6 t-Distributed Stochastic Neighbor Embedding (t-SNE) visualisation of 14 echo views from the 2-cell-DARTS model (128×128 image size). Each point represents an echo image from the test dataset, and different colored points represent different echo view classes.

Having fewer trainable parameters, both DARTS models also exhibit faster convergence and shorter training time per epoch than standard deeper network architectures: 157±116s vs. 622±576s, respectively, for the training dataset we used.

The confusion matrix for the 2-cell-DARTS model and image resolution of 128×128 pixels is provided in Figure 5. The errors appear predominantly clustered between a certain pair of views which represent anatomically adjacent imaging planes. The A5CH view proves to be the hardest one to detect (accuracy of about 80%), as the network is confused between this view and other apical windows. This is in line with previous observations that the greatest challenge lies in distinguishing between the various apical views.[31]

Interestingly, the two views the model found most difficult to correctly differentiate (A4CH-LV versus A5CH, and A2CH versus A3CH) were also the two views on which the two experts

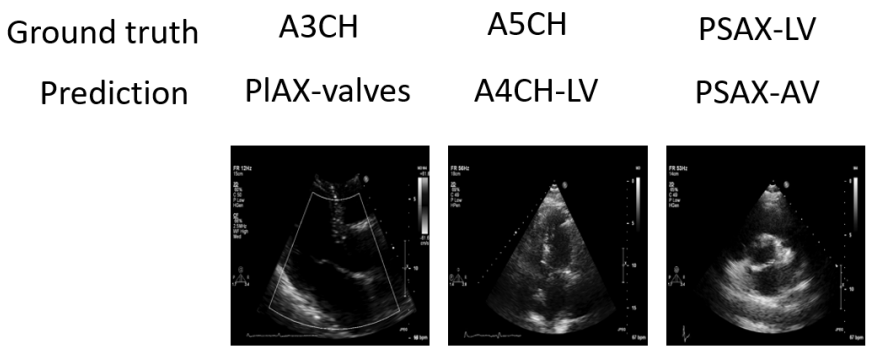

Fig 7 Three different misclassified examples predicted by the 2-cell-DARTS model for the image resolution of 128×128 pixels.

disagreed most often.[38] The A4CH view is in an anatomical continuity with the A5CH view. The difference is whether the scanning plane has been tilted to bring the aortic valve into view, which would make it A5CH. When the valve is only partially in view, or only in view during part of the cardiac cycle, the decision becomes a judgement call and there is room for disagreement. Similarly, the A3CH view differs from the A2CH view only in a rotation of the probe anticlockwise, again to bring the aortic valve into view

It is also interesting to note that the misclassification is not fully asymmetrical. For instance, while 42 cases of A5CH images are confused with A4CH-LV, there are only 14 occasions of A4CH-LV images mistaken for A5CH.

On the other hand, echo views with distinct characteristics are easier for the model to distinguish. For instance, PLAX-full and Suprasternal seem to have higher rates of correct identification, and the network is confused only on one occasion between these two views.

This is also evident on the t-Distributed Stochastic Neighbor Embedding (t-SNE) plot in Figure 6, which displays a planar representation of the internal high-dimensional organization of the 14 trained echo view classes within the network's final hidden layer (i.e. input data of the fully connected layer). Each point in the t-SNE plot represents an echo image from the test dataset.

Noticeably, not only has the network grouped similar images together (a cluster for each view, displayed with different color), but it has also grouped similar views together (highlighted with a unique background color). For instance, it has placed A5CH (blue) next to A4CH (dark brown), and indeed there is some "interdigitation" of such cases, e.g. for those whose classification between A4CH and A5CH might be debatable. Similarly, at the top right, the network has discovered that the features of the Subcostal-IVC images (green) are similar to the Subcostal images (red). This shows that the network can point to relationships and organizational patterns efficiently.

Figure 7 shows examples of misclassified cases, when the prediction of the 2-cell-DARTS model disagreed with the expert annotation. The error can be explained by the inherent difficulty of deciding, even for cardiologist experts, between views that are similar in appearance to human eyes and are in spatial continuity (case of A4CH / A5CH mix-up), images of poor quality (case of A4CH / PSAX mix-up), or views in which a same view-defining structure may be present (case of PSAX-LV / PSAX/AV mix-up).

4.3 Impact of image resolution, quality, and dataset size

The models seem to exhibit a plateau of accuracy between the two larger image resolutions of 96×96 and 128×128 pixels (Fig 8). On the other hand, for the smaller image size of 32×32 pixels, the classification performance seems to suffer across all network models, with a 2.3-5.1% reduction in accuracy relative to the resolution of 96×96 pixels.

Shown in Figure 9's upper panel, is the class-wise view detection accuracy for various input image resolutions. Notably, not all echo views are affected similarly by using lower image resolutions. The drop in overall performance is therefore predominantly caused by a marked decrease in detection accuracy of only certain views. For instance, A4CH-RV suffers a sharp reduction of

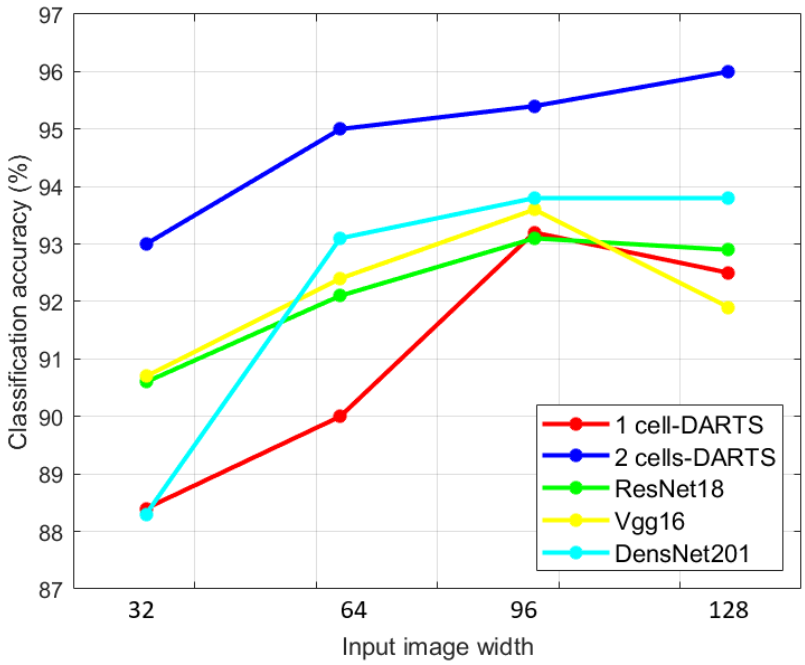

Fig 8 Comparison of accuracy for different classification models and different image resolutions; image width of 32 correspond to the image resolution of 32×32 pixels.

$>10\%$ in prediction accuracy when dealing with images of 32×32 pixels.

Figure 9's lower panel shows the relative confusion matrix, illustrating the improvement associated with using image resolution of 96×96 versus 32×32 pixels. Being already a difficult view to detect even in higher resolution images, A5CH will have 47 more cases of misclassified images when using images of 32×32 pixels. Overall, apical views seem to suffer the most from lower resolution images, being mainly misclassified as other apical views. For instance, the two classes associated with the A4CH will primarily be mistaken for one another. This is likely because, with a decreased resolution, the details of their distinct features would be less discernible by the network. Conversely, parasternal views seem to be less affected, and still detectable in downsampled images. This could be owing to the fact that the relevant features, on which the model relies for identifying this view, are still present and visible to the model.

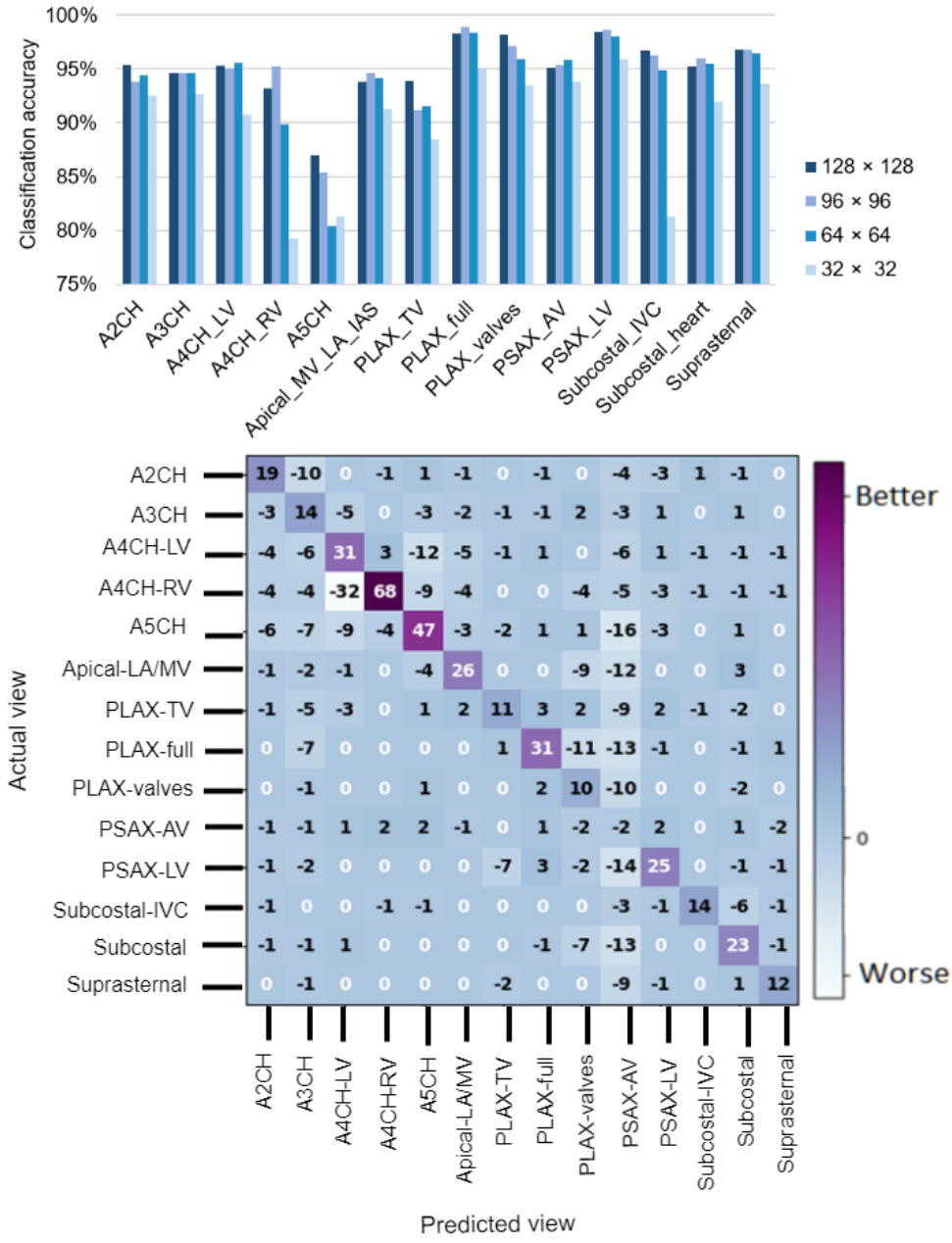

Fig 9 Accuracy of the 2-cell-DARTS model for various input image resolutions. Upper: class-wise prediction accuracy. Lower: relative confusion matrix showing improvement associated with using image resolution of 96×96 versus 32×32 pixels.

Overall, and for almost all echo views, the image size of 96×96 pixels appeared to be a good compromise between classification accuracy and computational costs.

To examine the influence of the size of the training dataset on the model's performance, we

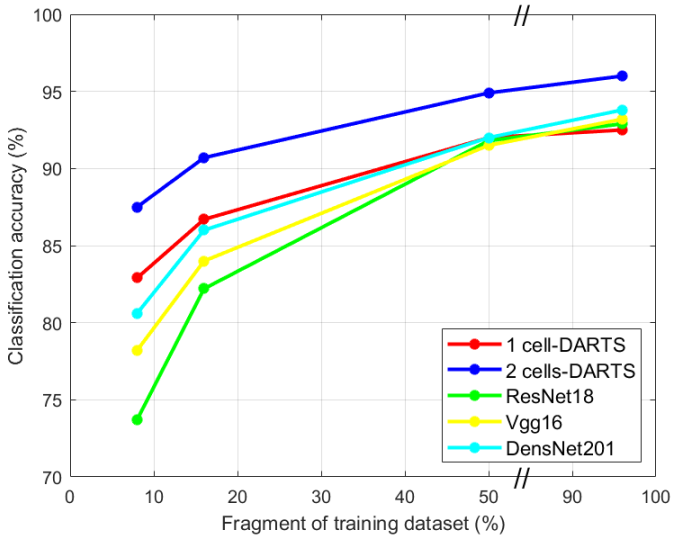

Fig 10 Comparison of accuracy of different classification models for image size of 128×128 versus different fragments of training dataset used when training the models. For each sub-dataset, all models were retrained from scratch.

conducted an additional experiment where we split the training data into sub-datasets with strict inclusion relationship (i.e., having the current sub-dataset a strict subset of the next sub-dataset), and ensured all the sub-datasets were consistent (i.e., having the same ratio for each echo view as in the original training dataset). We then retrained all targeted neural networks on these sub-datasets from scratch, and investigated how their accuracy varied with respect to the size of the dataset used for training the model. The size of the validation and testing datasets, however, remained unchanged.

Figure 10 shows a drop in the classification accuracy across all models when smaller sizes of training data are used for training the networks. However, various models are impacted differently. Suffering from redundancy, deeper neural networks require more training data to achieve similar performances. DenseNet, with the largest number of trainable parameters, appears to be the one which suffers the most, with a 20% reduction in its classification accuracy, when only 8% of the training dataset is used.

However, the DARTS-based models appear to be relatively less profoundly affected by the size of the training dataset, where both models demonstrate no more than 8% drop in their prediction accuracy when deprived of the full training dataset. When using fewer than 12,400 images (i.e., 50% of the training dataset), both DARTS-based models exhibit superior performance over the deeper networks.

Additionally, we hypothesised that the more numerous the echo view classes, the more difficult the task of distinguishing the views for deep learning models, e.g. because of more chances of misclassifications among classes. This is potentially the underlying reason for the inconsistent accuracies (84-97%) reported in the literature when classifying between 6 to 12 different view classes. To investigate this premise, we considered cases when only 5 or 7 different echo views were present in the dataset. To this end, rather than reducing the number of classes by merging several views to create new classes which may not be clinically very helpful, we were selective in choosing some of the existing classes. For each study, we aimed at including views representing anatomically adjacent or similar imaging planes such as apical windows (thus challenging for the models to distinguish), as well as other echo windows. The list of echo views included in each study is provided in Table 2.

The results show an increase in the overall prediction accuracy for the two DARTS-based models, when given the task of detecting fewer echo view classes and despite having relatively smaller training datasets to learn from. The 1-cell-DARTS model shows 8% improvement in its performance when the number of echo views is reduced from 14 to 5. The 2-cell-DARTS model reaches a maximum accuracy of 99.3%, i.e. higher than any previously reported accuracies for echo view classification. This highlights the fact that for a direct comparison of the classification accuracy between the models reported in literature, the number of different echo windows included

Table 2 The dependence of overall accuracy on the number of echo views; experimental results on the test dataset with 5, 7, and 14 classes for different network topologies, and image resolution of 64×64 pixels. The 7-class study included A2CH, A3CH, A4CH-LV, A5CH, PLAX-full, PSAX-LV, Subcostal-IVC, and a total of 24464 images. The 5-class study included A4CH-LV, PLAX-full, PSAX-AV, Subcostal, Suprasternal, and a total of 18896 images. Accuracy is ratio of correctly classified images to the total number of images; precision and recall are the macro average measures (average overall views of per-view measures); F1 score is the harmonic mean of precision and recall.

Network	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Parameters (thousands)	Inference Time (ms)	Time/epoch (s)
1-cell-DARTS							
14-classes	90.0	89.4	88.7	89.0	92	**6.5**	81
7-classes	96.4	96.1	96.1	96.1	110	7.8	58
5-classes	**98.1**	**98.3**	**97.9**	**98.1**	**85**	6.6	**38**
2-cell-DARTS							
14-classes	95.0	94.7	94.2	94.4	567	**12.6**	121
7-classes	97.0	96.9	96.7	96.8	709	15.6	85
5-classes	**99.3**	**99.3**	**99.1**	**99.2**	**556**	12.9	**55**

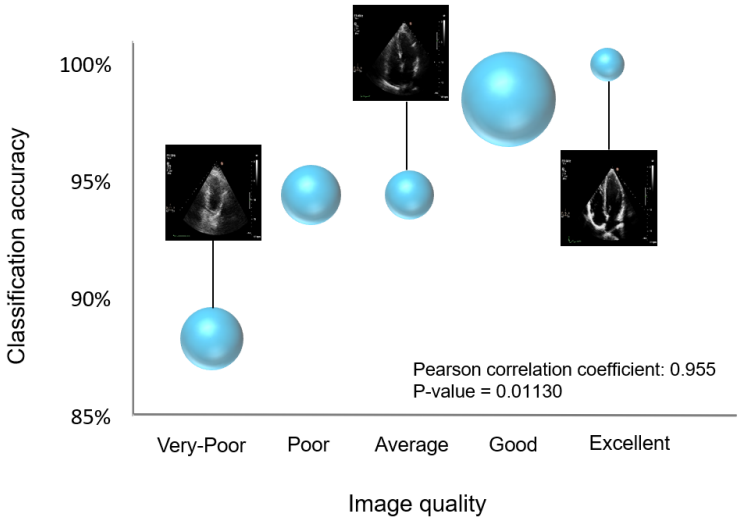

Fig 11 Correlation between the classification accuracy and the image quality (judged by the expert cardiologist) of A4CH-LV view in the test dataset. Area of the bubbles represent the relative frequency of the images in that quality score category. Results correspond to the the 2-cell-DARTS model and image resolution of 128×128 pixels. Here, p-value is the probability that the null hypothesis is true; i.e., the probability that the correlation between image quality and classification accuracy in the sample data occurred by chance.

in the study must be taken into account.

Finally, in order to study the impact of image quality on the classification performance, we

asked a second expert cardiologist to provide an assessment of image quality in the A4CH-LV views, and assign a quality label to each image where the quality was classified into 5 grades: very poor, poor, average, good, and excellent. Figure 11 displays the relationship between the classification accuracy of the 2-cell-DARTS model and the image quality in the test dataset. The area of the bubbles represents the relative frequency of the images in that quality score category, with the ”good” category as the dominant grade. This is likely because the image acquisition had been performed mainly by experienced echocardiographers.

The correlation between the classification accuracy and the image quality is evident (p-value of 0.01). Images labelled as having ”excellent” quality, indicated the highest classification accuracy of $\sim$100%. It is apparent that the discrepancy between the model's prediction and the expert annotation is higher in poor quality images. This could potentially be due to the fact that poorly visible chambers with a low degree of endocardial border delineation could result in some views being mistaken for other apical windows.

4.4 Study limitations and future work

This study sheds light on several possible directions for future work. Herein, we have focused on the rapid and accurate classification of individual frames from an echo cine loop. Such a task will be crucial for a real-time view detection system in clinical scenarios where images need to be processed while they are acquired from the patient and/or where the system is to be used for operator guidance. However, for offline studies and when the entire cine loop is available, classification of the echo videos could also be of practical use. Some studies have attempted video classification using the majority vote on some or all frames from a given video.[6,34] However, this approach does not use the temporal information available in the cine loop, such as the movement of structures

during the cardiac cycle. Therefore, a future study could look into using all available information for view detection.

Our study investigated 2D echocardiography as the clinically relevant modality. Currently, 3D echocardiography suffers from a considerable reduction in frame rate and image quality, and this has limited its adoption into routine practice over the past decade.[46–48] When such issues are resolved, automatic processing of the 3D modality could also be explored. In the meantime, 2D echocardiography remains unrivalled, particularly when high frame rates are needed.

We investigated the impact of image quality on the classification accuracy for apical four-chamber views only. A more comprehensive examination of the image quality and its influence on the detection of different echo views would be informative.

The dataset used in this study was comprised of images acquired using ultrasound equipment from GE and Philips manufacturers. Although the proposed models do not make any *a priori* assumptions on data obtained from specific vendors and therefore should be vendor-neutral, echo studies using more diverse ultrasound equipment should still be explored.

Similar to all previous studies, our dataset originated from one medical centre, i.e. Imperial College Healthcare NHS Trust's echocardiogram database. Representative multi-centre patient data will be essential for ensuring that the developed models will scale up well to other sites and environments.

Interpreting the results of the proposed models alongside other proposed architectures in the literature (with a wide range of reported accuracies) was not feasible. This is due to the fact that a direct comparison of the classification accuracy would require access to the same patient dataset. At present, no echocardiography dataset and corresponding annotations for view detection are publicly available.

In order to address such broadly acknowledged shortcomings in the application of deep learning to echocardiography, we are now developing Unity (data.unityimaging.net), a UK collaborative of cardiologists, physiologists and computer scientists, under the aegis of the British Society of Echocardiography. An image analysis interface has been developed in the form of a web-based, interactive, real-time platform to capture carefully-curated expert annotations from numerous echo specialists, with patient data provided by over a dozen sites across the UK, thus ensuring coverage of multiple vendors, systems and environments. All developed models designed using this annotation biobank (e.g., automated cardiac phase detection,[49] left ventricular segmentation,[50] and view classification in current study), will be made available under open-source agreements on intsav.github.io.

5 Conclusion

In this study, efficient CNN architectures are proposed for automated identification of the 2D echocardiographic views. The DARTS method was used in designing optimized architectures for rapid inference while maintaining high accuracy. A dataset of 14 different echocardiographic views was used for training and testing the proposed models. Compared with the standard classification CNN architectures, the proposed models are faster and achieve comparable classification performance. Such models can thus be used for real-time detection of the standard echo views.

The impact of image quality and size of the training dataset on the efficacy of the models was also investigated. Deeper neural network models, with a large number of redundant trainable parameters, require more training data to achieve similar performances. A direct correlation between the image quality of classification accuracy was observed.

The number of different echo views to be detected has a direct impact on the performance of

the deep learning models, and must be taken into account for a fair comparison of classification models.

Aggressively downsampled images will result in losing relevant features, thus lowering the prediction accuracy. On the other hand, while much larger images may be favoured for some fine grained applications (e.g., segmentation), their use for echo view classification would offer only slight improvements in performance (if any) at the expense of more processing and memory requirements.

Disclosures

No conflicts of interest are declared by the authors.

Acknowledgments

This work was supported in part by the British Heart Foundation (Grant no. PG/19/78/34733). N. Azarmehr is supported by the School of Computer Science, PhD scholarship at University of Lincoln, UK. We would like to express our gratitude to Piotr Bialecki for his valuable suggestions. We also thank Apostolos Vrettos for providing the expert annotations used to assess the impact of image quality.

References

1 R. M. Lang, L. P. Badano, V. Mor-Avi, *et al.*, "Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging," *European Heart Journal-Cardiovascular Imaging* **16**(3), 233–271 (2015).

2 H. Khamis, G. Zurakhov, V. Azar, *et al.*, "Automatic apical view classification of echocardiograms using a discriminative learning dictionary," *Medical Image Analysis* **36**, 15–21 (2017).

3 J. Zhang, S. Gajjala, P. Agrawal, *et al.*, "Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy," *Circulation* **138**(16), 1623–1635 (2018).

4 J. H. Park, S. K. Zhou, C. Simopoulos, *et al.*, "Automatic cardiac view classification of echocardiogram," in *2007 IEEE 11th International Conference on Computer Vision*, 1–8, IEEE (2007).

5 K. Siegersma, T. Leiner, D. Chew, *et al.*, "Artificial intelligence in cardiovascular imaging: state of the art and implications for the imaging cardiologist.," *Netherlands Heart Journal: Monthly Journal of the Netherlands Society of Cardiology and the Netherlands Heart Foundation* **27**(9), 403–413 (2019).

6 A. Østvik, E. Smistad, S. A. Aase, *et al.*, "Real-time standard view classification in transthoracic echocardiography using convolutional neural networks," *Ultrasound in medicine & biology* **45**(2), 374–384 (2019).

7 S. K. Zhou, J. Park, B. Georgescu, *et al.*, "Image-based multiclass boosting and echocardiographic view classification," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, **2**, 1559–1565, IEEE (2006).

8 J. Stoitsis, I. Valavanis, S. G. Mougiakakou, *et al.*, "Computer aided diagnosis based on medical image processing and artificial intelligence methods," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **569**(2), 591–595 (2006).

9 K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," *Computerized medical imaging and graphics* **31**(4-5), 198–211 (2007).

10 A. Coates, B. Huval, T. Wang, *et al.*, "Deep learning with cots hpc systems," in *International conference on machine learning*, 1337–1345 (2013).

11 M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and the future," *Classification in BioApps* , 323–350 (2018).

12 A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 1097–1105 (2012).

13 G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, "A survey on deep learning in medical image analysis," *Medical image analysis* **42**, 60–88 (2017).

14 B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, OpenReview.net (2017).

15 H. Pham, M. Guan, B. Zoph, *et al.*, "Efficient neural architecture search via parameters sharing," in *International Conference on Machine Learning*, 4095–4104, PMLR (2018).

16 H. Liu, K. Simonyan, and Y. Yang, "Darts: Differentiable architecture search," in *International Conference on Learning Representations*, (2018).

17 S. Xie, H. Zheng, C. Liu, *et al.*, "SNAS: stochastic neural architecture search," in *International Conference on Learning Representations*, (2019).

18 F. Hutter, L. Kotthoff, and J. Vanschoren, *Automated machine learning: methods, systems, challenges*, Springer Nature (2019).

19 I. Bello, B. Zoph, V. Vasudevan, *et al.*, "Neural optimizer search with reinforcement learning," in *International Conference on Machine Learning*, 459–468, PMLR (2017).

20 B. Zoph, V. Vasudevan, J. Shlens, *et al.*, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8697–8710 (2018).

21 E. Real, A. Aggarwal, Y. Huang, *et al.*, "Regularized evolution for image classifier architecture search," in *Proceedings of the aaai conference on artificial intelligence*, **33**, 4780–4789 (2019).

22 E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650 (2019).

23 S. Ebadollahi, S.-F. Chang, and H. Wu, "Automatic view recognition in echocardiogram videos using parts-based representation," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, **2**, II–II, IEEE (2004).

24 D. Agarwal, K. Shriram, and N. Subramanian, "Automatic view classification of echocardiograms using histogram of oriented gradients," in *2013 IEEE 10th International Symposium on Biomedical Imaging*, 1368–1371, IEEE (2013).

25 H. Wu, D. M. Bowers, T. T. Huynh, *et al.*, "Echocardiogram view classification using low-level features," in *2013 IEEE 10th International Symposium on Biomedical Imaging*, 752–755, IEEE (2013).

26 R. Kumar, F. Wang, D. Beymer, *et al.*, "Cardiac disease detection from echocardiogram using

edge filtered scale-invariant motion features," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, 162–169, IEEE (2010).

27 M. Otey, J. Bi, S. Krishna, *et al.*, "Automatic view recognition for cardiac ultrasound images," in *International Workshop on Computer Vision for Intravascular and Intracardiac Imaging*, 187–194 (2006).

28 D. Beymer, T. Syeda-Mahmood, and F. Wang, "Exploiting spatio-temporal information for view recognition in cardiac echo videos," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 1–8, IEEE (2008).

29 R. Kumar, F. Wang, D. Beymer, *et al.*, "Echocardiogram view classification using edge filtered scale-invariant motion features," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 723–730, IEEE (2009).

30 X. Gao, W. Li, M. Loomes, *et al.*, "A fused deep learning architecture for viewpoint classification of echocardiography," *Information Fusion* **36**, 103–113 (2017).

31 R. C. Deo, J. Zhang, L. A. Hallock, *et al.*, "An end-to-end computer vision pipeline for automated cardiac function assessment by echocardiography," *CoRR* (2017).

32 A. Madani, R. Arnaout, M. Mofrad, *et al.*, "Fast and accurate view classification of echocardiograms using deep learning," *NPJ digital medicine* **1**(1), 1–8 (2018).

33 O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 234–241, Springer (2015).

34 A. Madani, J. R. Ong, A. Tibrewal, *et al.*, "Deep echocardiography: data-efficient supervised

and semi-supervised deep learning towards automated diagnosis of cardiac disease," *NPJ digital medicine* **1**(1), 1–11 (2018).

35 C. Szegedy, V. Vanhoucke, S. Ioffe, *et al.*, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826 (2016).

36 G. Huang, Z. Liu, L. Van Der Maaten, *et al.*, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).

37 H. Vaseli, Z. Liao, A. H. Abdi, *et al.*, "Designing lightweight deep learning models for echocardiography view classification," in *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*, **10951**, 109510F, International Society for Optics and Photonics (2019).

38 J. P. Howard, J. Tan, M. J. Shun-Shin, *et al.*, "Improving ultrasound video classification: an evaluation of novel deep learning methods in echocardiography," *Journal of medical artificial intelligence* **3** (2020).

39 K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, (2015).

40 K. He, X. Zhang, S. Ren, *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

41 E. Real, S. Moore, A. Selle, *et al.*, "Large-scale evolution of image classifiers," in *International Conference on Machine Learning*, 2902–2911, PMLR (2017).

631 42 M. Botvinick, S. Ritter, J. X. Wang, *et al.*, "Reinforcement learning, fast and slow," *Trends in*
632 *cognitive sciences* **23**(5), 408–422 (2019).
633 43 T. Elsken, J. H. Metzen, F. Hutter, *et al.*, "Neural architecture search: A survey.," *J. Mach.*
634 *Learn. Res.* **20**(55), 1–21 (2019).
635 44 D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International*
636 *Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015,*
637 *Conference Track Proceedings*, (2015).
638 45 A. Paszke, S. Gross, S. Chintala, *et al.*, "Automatic differentiation in pytorch," (2017).
639 46 K. Cheng, M. Monaghan, A. Kenny, *et al.*, "3d echocardiography: Benefits and steps to wider
640 implementation," *Br J Cardiol* **25**, 63–68 (2018).
641 47 D. Loeckx, J. Ector, F. Maes, *et al.*, "Spatiotemporal non-rigid image registration for 3d
642 ultrasound cardiac motion estimation," in *Medical Imaging 2007: Ultrasonic Imaging and*
643 *Signal Processing*, **6513**, 65130X, International Society for Optics and Photonics (2007).
644 48 P. Carnahan, J. Moore, D. Bainbridge, *et al.*, "Multi-view 3d echocardiography volume com-
645 pounding for mitral valve procedure planning," in *Medical Imaging 2020: Image-Guided*
646 *Procedures, Robotic Interventions, and Modeling*, **11315**, 1131510, International Society for
647 Optics and Photonics (2020).
648 49 E. S. Lane, N. Azarmehr, J. Jevsikov, *et al.*, "Multibeat echocardiographic phase detection
649 using deep neural networks," *Computers in Biology and Medicine* , 104373 (2021).
650 50 N. Azarmehr, X. Ye, F. Janan, *et al.*, "Automated segmentation of left ventricle in 2d echocar-
651 diography using deep learning," in *International Conference on Medical Imaging with Deep*
652 *Learning – Extended Abstract Track*, (London, United Kingdom) (2019).

Neda Azarmehr is a Research Associate at the University of Sheffield. In May 2017, Neda was awarded a PhD scholarship at the University of Lincoln, in collaboration with Imperial College London to develop automated models using deep learning, computer vision algorithm to assess the left ventricle function which enables physicians to analyse cardiac echo images more precisely. Her research focuses on developing models using Artificial Intelligence, Deep Learning, Computer Vision to support clinicians in decision-making.

Xujiong Ye is a Professor of Medical Imaging and Computer Vision in the School of Computer Science, University of Lincoln, UK. Prof. Ye has over 20 years of research and development experience in medical imaging and computer vision from both academia and industry. Her main research is to develop computational models using advanced image analysis, computer vision, and artificial intelligence to support clinicians in decision-making.

James P Howard is currently undertaking his PhD, "Machine Learning in Cardiovascular Imaging", at Imperial College London, UK. His research interests include the applications of constitutional neural networks in the processing of echocardiograms, cardiac magnetic resonance imaging, and coronary physiology waveform analysis.

Elisabeth Sarah Lane completed an MSc in Software Engineering in 2019 and is currently a PhD candidate at The University of West London. Her current research focus is the application of Deep Learning algorithms for automatic phase detection in echocardiograms. Beth's interests include Machine Learning, Computer Vision and Artificial Intelligence for the analysis and interpretation of clinical imaging.

Robert Labs is currently undertaking his PhD, Artificial Intelligence for automated quality

assessment of 2D echocardiography at University of West London, UK. His research interests include the clinical application of machine learning for accurate quantification and diagnosis of cardiac infarction.

Matthew J Shun-shin is a Clinical Lecturer in Cardiology at National Heart and Lung Institute, Imperial College London. His research interests include Artificial Intelligence, Echocardiography, valvular heart disease, and improving device therapies for heart failure.

Graham D Cole is a Consultant Cardiologist primarily based at Hammersmith Hospital, part of Imperial College Healthcare NHS Trust. Graham qualified from Gonville and Caius College, University of Cambridge in 2005. He subsequently completed a four-year fellowship in cardiac MRI at Heart Hospital Imaging Centre, the London CT fellowship and a PhD in echocardiography. He has interests in the optimal use of cardiac imaging and research reliability.

Luc Bidaut has worked with and on most aspects of biomedical imaging and technology in highly multidisciplinary research, clinical and translational international environments, always in direct collaboration with all relevant stakeholders from scientific, technical and medical disciplines. His active involvement in the development, implementation and actual deployment of related technologies and applications was and remains primarily focused on maximizing the utility and actionability of the information collected through imaging modalities and other sensors, at various stages of the translational pipeline or clinical workflow.

Darrel P Francis is a Professor of Cardiology at the National Heart and Lung Institute, Imperial College London. He specialises in using quantitative techniques, derived from mathematics, engineering, and statistics, to problems that affect patients with heart disease.

Massoud Zolgharni is a Professor of Computer Vision at the School of Computing and Engineering, University of West London. He is also an Honorary Research Fellow at the National Heart and Lung Institute, Imperial College London. His research interests include Computer Vision, Medical Imaging, Machine Learning, and Numerical Simulations.

List of Figures

4 Optimal normal and reduction cells for the input image size of 128×128 pixels, as suggested by the DARTS method, where 3×3 and 5×5 dilated separable convolutions, 3×3 max-pooling, and skip-connection operations have been retained from the candidate operations initially included. Each cell has 2 inputs which are the cell outputs in the previous two layers. The output of the cell is defined as the depth-wise concatenation of all nodes in the cell. A schematic view of the "2-cell-DARTS", formed from a sequential stack of 2 cells, is also displayed on the left. Stem layer incorporates a convolution layer and a batch normalisation layer.

5 Confusion matrix for the 2-cell-DARTS model and input image resolution of 128×128 pixels.

6 t-Distributed Stochastic Neighbor Embedding (t-SNE) visualisation of 14 echo views from the 2-cell-DARTS model (128×128 image size). Each point represents an echo image from the test dataset, and different colored points represent different echo view classes.

7 Three different misclassified examples predicted by the 2-cell-DARTS model for the image resolution of 128×128 pixels.

8 Comparison of accuracy for different classification models and different image resolutions; image width of 32 correspond to the image resolution of 32×32 pixels.

9 Accuracy of the 2-cell-DARTS model for various input image resolutions. Upper: class-wise prediction accuracy. Lower: relative confusion matrix showing improvement associated with using image resolution of 96×96 versus 32×32 pixels.

List of Tables

2 The dependence of overall accuracy on the number of echo views; experimental results on the test dataset with 5, 7, and 14 classes for different network topologies, and image resolution of 64×64 pixels. The 7-class study included A2CH, A3CH, A4CH-LV, A5CH, PLAX-full, PSAX-LV, Subcostal-IVC, and a total of 24464 images. The 5-class study included A4CH-LV, PLAX-full, PSAX-AV, Subcostal, Suprasternal, and a total of 18896 images. Accuracy is ratio of correctly classified images to the total number of images; precision and recall are the macro average measures (average overall views of per-view measures); F1 score is the harmonic mean of precision and recall.